Concentration et inégalités de marge

Olivier Catoni

3 juin 2013

1. Déviations des sommes de variables aléatoires indépendantes

Soit X_i, $1 \leq i \leq n$ un échantillon de variables indépendantes et

$$M \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} X_i$$

leur moyenne empirique. On se pose la question des déviations de M par rapport à sa moyenne

$$m \overset{\text{def}}{=} \mathbb{E}(M) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i).$$

Considérons les fonctions génératrices des moments

$$\psi_i(\lambda) = \log\Bigl\{\mathbb{E}\bigl[\exp(\lambda X_i)\bigr]\Bigr\},$$
$$\psi(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \psi_i(\lambda).$$

Ce sont les fonctions convexes, nulles en zéro, à valeurs dans $\mathbb{R}_+ \cup \{+\infty\}$. Considérons la fonction duale

$$\psi^*(x) = \sup_{\lambda \in \mathbb{R}_+} \lambda x - \psi(\lambda) \in \mathbb{R}_+ \cup \{+\infty\}.$$

Proposition 1.1 *Les déviations de la moyennes empirique M vérifient*

$$\mathbb{P}\bigl(M \geq x\bigr) \leq \exp\bigl[-n\psi^*(x)\bigr].$$

Preuve.

CNRS – UMR 8553, Département de Mathématiques et Applications, Ecole Normale Supérieure, 45, rue d'Ulm, F75230 Paris cedex 05, and INRIA Paris-Rocquencourt – CLASSIC team.

$$\begin{aligned}\mathbb{P}\bigl(M \geq x\bigr) = \mathbb{E}\Bigl\{\mathbb{1}\bigl[\exp\bigl(n\lambda(M-x)\bigr) \geq 1\bigr]\Bigr\} \\ \leq \mathbb{E}\bigl[\exp\bigl(n\lambda(M-x)\bigr)\bigr] \\ = \exp\bigl\{n\bigl[\psi(\lambda) - \lambda x\bigr]\bigr\}, \qquad \lambda \in \mathbb{R}_+.\end{aligned}$$

Par conséquent

$$\mathbb{P}\bigl(M \geq x\bigr) \leq \inf_{\lambda\in\mathbb{R}_+} \exp\bigl\{n\bigl[\psi(\lambda) - \lambda x\bigr]\bigr\} = \exp\bigl(-n\psi^*(x)\bigr).$$

□

PROPOSITION 1.2 *Posons* $\Lambda_i = \sup\bigl\{\lambda \in \mathbb{R}_+ : \psi_i(\lambda) < +\infty\bigr\}$ *et* $\Lambda = \min\bigl\{\Lambda_1, \dots, \Lambda_n\bigr\}$. *Pour tout* $\lambda \in [0, \Lambda_i[$, $\psi_i(\lambda) < +\infty$ *et la fonction* ψ_i *est de classe* $\mathcal{C}^\infty$ *sur* $]0, \Lambda_i[$ *lorsque* $\Lambda_i > 0$. *Si de plus* $\mathbb{E}\bigl(|X_i|^k\bigr) < \infty$, *la fonction* ψ_i *est de classe* $\mathcal{C}^k$ *sur* $[0, \Lambda_i[$.

PREUVE. Posons $\varphi(\lambda) = \mathbb{E}\bigl[\exp(\lambda X_i)\bigr]$. Soit $\lambda \in [0, \Lambda_i[$. Par définition de Λ_i, il existe $\beta \in]\lambda, \Lambda_i[$ tel que $\psi_i(\beta) < \infty$, et donc $\varphi(\beta) < \infty$. D'après l'inégalité de Jensen

$$+\infty > \mathbb{E}\bigl[\exp(\beta X_i)\bigr] = \mathbb{E}\Bigl\{\bigl[\exp(\lambda X_i)\bigr]^{\beta/\lambda}\Bigr\} \geq \Bigl\{\mathbb{E}\bigl[\exp(\lambda X_i)\bigr]\Bigr\}^{\beta/\lambda},$$

ce qui prouve que $\varphi_i(\lambda) < \infty$, et donc que $\psi_i(\lambda) < \infty$.

Remarquons que

$$X_i^{j-1}\exp(\beta X_i) = X_i^{j-1}\exp(\alpha X_i) + \int_\alpha^\beta X_i^j \exp(\lambda X_i)\,\mathrm{d}\lambda, \qquad 0 < \alpha < \beta < \Lambda_i, \quad j \geq 1.$$

De plus

$$\mathbb{E}\Bigl\{\sup_{\lambda\in[\alpha,\beta]}\bigl|X_i^j \exp\bigl(\lambda X_i\bigr)\bigr|\Bigr\} < \infty$$

En effet, considérons $\gamma \in]\beta, \Lambda_i[$ et

$$\begin{aligned}C_1 &= \sup_{x\in\mathbb{R}_+} x^j \exp\bigl[-(\gamma-\beta)x\bigr],\\ C_2 &= \sup_{x\in\mathbb{R}_+} x^j \exp(-\alpha x).\end{aligned}$$

$$\bigl|X_i^j \exp(\lambda X_i)\bigr| \leq \begin{cases} C_1 \exp(\gamma X_i), & X_i \geq 0,\\ C_2, & X_i \leq 0.\end{cases}$$

Par conséquent

$$\mathbb{E}\Bigl\{\sup_{\lambda\in[\alpha,\beta]}\bigl|X_i^j\exp(\lambda X_i)\bigr|\Bigr\}\leq C_1\mathbb{E}\bigl[\exp(\gamma X_i)\bigr]+C_2<\infty.$$

On peut donc employer le théorème de Fubini et écrire

$$\begin{aligned}\mathbb{E}\bigl[X_i^{j-1}\exp(\beta X)\bigr]&=\mathbb{E}\bigl[X_i^{j-1}\exp(\alpha X_i)\bigr]+\mathbb{E}\biggl(\int_\alpha^\beta X_i^j\exp(\lambda X_i)\,\mathrm{d}\lambda\biggr)\\&=\mathbb{E}\bigl[X_i^{j-1}\exp(\alpha X_i)\bigr]+\int_\alpha^\beta\mathbb{E}\bigl[X_i^j\exp(\lambda X_i)\bigr]\,\mathrm{d}\lambda.\end{aligned}$$

D'après le théorème de convergence dominée, $\lambda\mapsto\mathbb{E}\bigl(X_i^j\exp(\lambda X_i)\bigr]:[\alpha,\beta]\to\mathbb{R}$ est continue, donc $\beta\mapsto\mathbb{E}\bigl[X_i^{j-1}\exp(\beta X_i)\bigr]$ est de classe $\mathcal{C}^1$, de dérivée $\mathbb{E}\bigl[X_i^j\exp(\beta X_j)\bigr]$, donc $\beta\mapsto\mathbb{E}\bigl[\exp(\beta X_i)\bigr]$ est de classe $\mathcal{C}^\infty$ sur $]0,\Lambda_i[$ et il en est de même de ψ_i.

Supposons maintenant de plus que $\mathbb{E}\bigl[|X_i|^k\bigr]<\infty$. On peut dans ce cas montrer de façon analogue que

$$\mathbb{E}\Bigl\{\sup_{\lambda\in[0,\beta]}\bigl|X_i^j\exp(\lambda X_i)\bigr|\Bigr\}<\infty,\qquad 0<\beta<\Lambda_i,$$

en déduire que

$$\mathbb{E}\bigl[X_i^{j-1}\exp(\beta X_i)\bigr]=\mathbb{E}\bigl(X_i^{j-1}\bigr)+\int_0^\beta\mathbb{E}\bigl[X_i^j\exp(\lambda X_i)\bigr]\,\mathrm{d}\lambda,\qquad 0\leq\beta<\Lambda_i,\quad 1\leq j\leq k,$$

et enfin que φ_i et ψ_i sont de classe $\mathcal{C}^k$ sur $[0,\Lambda_i[$. □

Proposition 1.3 *Supposons que $\mathbb{E}\bigl(X_i^2\bigr)<\infty$ et que $\Lambda_i>0$. La dérivée seconde de ψ_i prend la forme d'une variance :*

$$\psi_i''(\lambda)=\frac{\mathbb{E}\bigl[X_i^2\exp(\lambda X_i)\bigr]}{\mathbb{E}\bigl[\exp(\lambda X_i)\bigr]}-\biggl(\frac{\mathbb{E}\bigl[X_i\exp(\lambda X_i)\bigr]}{\mathbb{E}\bigl[\exp(\lambda X_i)\bigr]}\biggr)^2,\qquad 0\leq\lambda<\Lambda_i,$$

de plus

$$\psi_i(\lambda)=\lambda\mathbb{E}(X_i)+\int_0^\lambda(\lambda-\alpha)\psi_i''(\alpha)\,\mathrm{d}\alpha,\qquad 0\leq\lambda<\Lambda_i.$$

Preuve. D'après la proposition précédente, ψ_i est de classe $\mathcal{C}^2$ sur $[0,\Lambda_i[$ et

$$\psi_i'(\lambda)=\frac{\mathbb{E}\bigl[X_i\exp(\lambda X_i)\bigr]}{\mathbb{E}\bigl[\exp(\lambda X_i)\bigr]},$$

en dérivant une fois de plus, on obtient l'expression de ψ_i'' indiquée dans la proposition. Considérons la variable aléatoire Y_i de loi

$$\mathbb{P}\bigl(Y_i \in A\bigr) = \frac{\mathbb{E}\bigl[\mathbb{1}(X_i \in A) \exp(\lambda X_i)\bigr]}{\mathbb{E}\bigl[\exp(\lambda X_i)\bigr]},$$

pour tout borélien A. Elle vérifie pour toute fonction mesurable f telle que $\mathbb{E}\bigl[|f(X_i)| \exp(\lambda X_i)\bigr] < \infty$

$$\mathbb{E}\bigl[f(Y_i)\bigr] = \frac{\mathbb{E}\bigl[f(X_i) \exp(\lambda X_i)\bigr]}{\mathbb{E}\bigl[\exp(\lambda X_i)\bigr]},$$

ce qui montre que $\psi_i''(\lambda) = \mathbb{E}\bigl(Y_i^2\bigr) - \mathbb{E}\bigl(Y_i\bigr)^2$ est bien une variance. La deuxième partie de la proposition s'obtient en écrivant la formule de Taylor $\psi_i(\lambda) = \psi_i(0) + \lambda\psi_i'(0) + \int_0^\lambda (\lambda - \alpha)\psi_i''(\alpha)\,\mathrm{d}\alpha$. □

Proposition 1.4 *Supposons que $\Lambda > 0$ et que $\mathbb{E}\bigl(X_i^2\bigr) < \infty$, $1 \leq i \leq n$. Posons*

$$\overline{V}(\lambda) \overset{\text{def}}{=} \frac{2}{\lambda^2}\bigl[\psi(\lambda) - \lambda m\bigr] = \frac{2}{\lambda^2}\int_0^\lambda (\lambda - \alpha)\psi''(\alpha)\,\mathrm{d}\alpha, \qquad 0 \leq \lambda < \Lambda$$

$$V(\lambda) \overset{\text{def}}{=} \sup_{\beta \in [0,\lambda]} \overline{V}(\beta),$$

$$v \overset{\text{def}}{=} V(0) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Bigl\{\bigl[X_i - \mathbb{E}(X_i)\bigr]^2\Bigr\}$$

et remarquons que V est une fonction continue et croissante à valeurs dans $\mathbb{R}+\cup\{+\infty\}$. Sous ces hypothèses

$$\mathbb{P}\bigl(M \geq m + x\bigr) \leq \exp\biggl(-\frac{nx^2}{2V(x/v)}\biggr),$$

$$\mathbb{P}\left(M \geq m + \sqrt{\frac{2\log(\epsilon^{-1})}{n} V\left(\sqrt{\frac{2\log(\epsilon^{-1})}{nv}}\right)}\right) \leq \epsilon.$$

Preuve. Pour tout $0 \leq \beta \leq \lambda$,

$$\psi^*(m + x) \geq \beta x - \frac{\beta^2}{2}V(\lambda),$$

si bien que

$$\mathbb{P}\bigl(M \geq m + x\bigr) \leq \exp\biggl[-n\biggl(\beta x - \frac{\beta^2}{2}V(\lambda)\biggr)\biggr].$$

On obtient la première inégalité en choisissant $\lambda = x/v$ et $\beta = x/V(\lambda) \leq \lambda$. Pour obtenir la seconde posons $\epsilon = \exp\Bigl[-n\Bigl(\beta x - \frac{\beta^2}{2}V(\lambda)\Bigr)\Bigr]$, pour obtenir dans un premier temps

$$\mathbb{P}\biggl(M \geq m + \frac{\beta}{2}V(\lambda) + \frac{\log(\epsilon^{-1})}{n\beta}\biggr) \leq \epsilon.$$

Choisissons alors $\lambda = \sqrt{\frac{2\log(\epsilon^{-1})}{nv}} \geq \beta = \sqrt{\frac{2\log(\epsilon^{-1})}{nV(\lambda)}}$ pour conclure. □

Proposition 1.5 (Inégalité de Bennett) *Supposons que* $\mathbb{E}\bigl(X_i^2\bigr) < \infty$ *et que* $X_i \leq \mathbb{E}(X_i) + b$, $1 \leq i \leq n$. *Introduisons la fonction*

$$h(u) = (1+u)\log(1+u) - u \geq \frac{u^2}{2(1+u/3)}, \qquad u \in \mathbb{R}_+.$$

Sous ces hypothèses

$$\mathbb{P}\bigl(M \geq m + x\bigr) \leq \exp\biggl[-\frac{nv}{b^2}h\Bigl(\frac{bx}{v}\Bigr)\biggr] \leq \exp\biggl(-\frac{nx^2}{2v + \frac{2bx}{3}}\biggr),$$
$$\mathbb{P}\left(M \geq m + \sqrt{\frac{2v\log(\epsilon^{-1})}{n}}\biggl(1 - \frac{b}{3v}\sqrt{\frac{2v\log(\epsilon^{-1}}{n}}\biggr)^{-1/2}\right) \leq \epsilon.$$

Preuve. Remarquons tout d'abord que pour tout $\lambda \in \mathbb{R}+$,

$$\psi^*(m+x) \geq \lambda(x+m) - \frac{1}{n}\sum_{i=1}^n \log\bigl[\mathbb{E}\bigl(\exp(\lambda X_i)\bigr)\bigr]$$
$$= \lambda x - \frac{1}{n}\sum_{i=1}^n \log\Bigl\{\mathbb{E}\bigl[\exp\bigl(\lambda(X_i - m_i)\bigr)\bigr]\Bigr\},$$

où $m_i \overset{\text{def}}{=} \mathbb{E}(X_i)$. On peut alors écrire

$$\mathbb{E}\bigl[\exp\bigl(\lambda(X_i - m_i)\bigr)\bigr] - 1 = \mathbb{E}\bigl[\exp\bigl(\lambda(X_i - m_i)\bigr) - 1 - \lambda(X_i - m_i)\bigr]$$
$$= \mathbb{E}\bigl[\lambda^2(X_i - m_i)^2 g\bigl(\lambda(X_i - m_i)\bigr)\bigr]$$

où $g(y) = y^{-2}\bigl(\exp(y) - 1 - y\bigr)$. La fonction g est croissante sur $\mathbb{R}$. En effectuant un développement de Taylor à l'ordre deux de la fonction $z \mapsto \exp(yz)$ entre 0 et 1, on voit en effet qu'elle peut s'écrire

$$g(y) = \int_0^1 (1-z)\exp(yz)\,\mathrm{d}z, \qquad y \in \mathbb{R}.$$

On en déduit que

$$\mathbb{E}\bigl[\lambda^2(X_i - m_i)^2 g\bigl(\lambda(X_i - m_i)\bigr)\bigr] \leq \mathbb{E}\bigl[\lambda^2(X_i - m_i)^2 g(\lambda b)\bigr], \qquad 1 \leq i \leq n,$$

et donc que

$$\log\Bigl\{\mathbb{E}\bigl[\exp\bigl(\lambda(X_i - m_i)\bigr)\bigr]\Bigr\} \leq \lambda^2 g(\lambda b)\mathbb{E}\bigl[(X_i - m_i)^2\bigr].$$

Ainsi

$$\psi^*(m + x) \geq \lambda x - \lambda^2 v g(\lambda b) = \lambda x - \frac{v}{b^2}\bigl(\exp(\lambda b) - 1 - \lambda b\bigr).$$

Choisissons $\lambda = b^{-1} \log\left(1 + \frac{bx}{v}\right)$ pour obtenir

$$\psi^*(x) \geq \frac{v}{b^2} h\left(\frac{bx}{v}\right).$$

Montrons maintenant que $h(u) \geq \frac{u^2}{2(1 + u/3)}$, $u > -1$. Calculons les dérivées de h, $h'(u) = \log(1 + u)$, $h''(u) = 1/(1 + u)$, puis les dérivées de $f(u) = (1 + u/3)h(u) - u^2/2$. On obtient $f'(u) = h'(u)(1 + u/3) + h(u)/3 - u$ qui vérifie $f'(0) = 0$ et

$$\begin{aligned} f''(u) = h''(u)(1 + u/3) + 2h'(u)/3 - 1 &= \frac{1 + u/3}{1 + u} + \frac{2}{3}\log(1 + u) - 1 \\ &= \frac{2}{3}\log(1 + u) - \frac{2u}{3(1 + u)} = \frac{2h(u)}{3(1 + u)} \geq 0, \qquad u > -1. \end{aligned}$$

La fonction f, convexe, nulle en zéro et de dérivée première nulle en zéro est donc positive.

Posons $\epsilon = \exp\left(-\frac{nx^2}{2v + \frac{2bx}{3}}\right)$. On obtient

$$\begin{aligned} x^2 &= \frac{2v\log(\epsilon^{-1})}{n}\left(1 + \frac{bx^2}{3vx}\right) \\ &\leq \frac{2v\log(\epsilon^{-1}}{n}\left(1 + \frac{bx^2}{3v}\left(\frac{2v\log(\epsilon^{-1})}{n}\right)^{-1/2}\right). \end{aligned}$$

On en déduit que

$$x^2 \leq \frac{2v\log(\epsilon^{-1})}{n}\left(1 - \frac{b}{3v}\sqrt{\frac{2v\log(\epsilon^{-1})}{n}}\right)^{-1},$$

ce qui prouve la deuxième inégalité de la proposition. $\square$

Proposition 1.6 (Inégalité de Hoeffding) *Supposons que $a_i \leq X_i \leq b_i$, $1 \leq i \leq n$. Dans ce cas*

$$\mathbb{P}\bigl(M \geq m + x\bigr) \leq \exp\biggl(-\frac{2n^2x^2}{\sum_{i=1}^n (b_i - a_i)^2}\biggr),$$
$$\mathbb{P}\left(M \geq m + \sqrt{\frac{\sum_{i=1}^n (b_i - a_i)^2 \log(\epsilon^{-1}}{2n^2}}\right) \leq \epsilon.$$

Preuve. La dérivée seconde de ψ_i est la variance d'une variable aléatoire à valeurs dans l'intervalle $[a_i, b_i]$, et ne peut donc excéder $(b_i - a_i)^2/4$. Par conséquent $\psi(\lambda) \leq \lambda m + \dfrac{\lambda^2}{8}\displaystyle\sum_{i=1}^n (b_i - a_i)^2$ d'où $\psi^*(m + x) \geq \dfrac{2nx^2}{\sum_{i=1}^n (b_i - a_i)^2}$.
□

2. Bornes PAC-Bayésiennes pour les déviations uniformes des moyennes empiriques par rapport à leurs espérances

Considérons n variables aléatoires indépendantes X_i, $1 \leq i \leq n$ à valeurs dans un espace mesurable $\mathcal{X}$, un espace mesurable de paramètres Θ et une fonction mesurable $f : \mathcal{X} \times \Theta \to \mathbb{R}$ (qui peut être vue comme une famille de fonctions de $\mathcal{X}$ dans $\mathbb{R}$ indexée par θ). Supposons que

$$\mathbb{E}\bigl[f(X,\theta)^2\bigr] < +\infty, \qquad \theta \in \Theta,$$

et posons

$$\begin{aligned} M(\theta) &= \frac{1}{n}\sum_{i=1}^n f(X_i,\theta), \\ m(\theta) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\bigl[f(X_i,\theta)\bigr], \\ \psi_i(\lambda,\theta) &= \log\Bigl\{\mathbb{E}\exp\bigl[\lambda f(X_i,\theta)\bigr]\Bigr\}, \\ \psi(\lambda,\theta) &= \frac{1}{n}\sum_{i=1}^n \psi_i(\lambda,\theta), \\ \Lambda &= \sup\bigl\{\lambda : \psi(\lambda,\theta) < \infty, \theta \in \Theta\bigr\} \end{aligned}$$

Proposition 2.1 *Supposons que $\Lambda > 0$. Soit $\nu \in \mathcal{M}_+^1(\Theta)$ une mesure de référence sur l'espace des paramètres Θ. Pour tout $\lambda \in [0, \Lambda[$,*

$$\mathbb{E}\Biggl[\exp\Biggl(\sup\biggl\{\int_\Theta n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\,\mathrm{d}\rho(\theta)-\mathcal{K}(\rho,\nu),$$
$$\rho\in\mathcal{M}_+^1(\Theta),\theta\mapsto\lambda M(\theta)-\psi(\lambda,\theta)\in\mathbb{L}^1(\rho),\mathcal{K}(\rho,\nu)<\infty\biggr\}\Biggr)\Biggr]\leq 1.$$

Par conséquent, avec probabilité au moins $1-\epsilon$, pour toute probabilité $\rho\in\mathcal{M}_+^1(\Theta)$, telle que $\theta\mapsto\lambda M(\theta)-\psi(\lambda,\theta)\in\mathbb{L}^1(\rho)$ et $\mathcal{K}(\rho,\nu)<\infty$,

$$\int M(\theta)\,\mathrm{d}\rho(\theta)\leq\frac{1}{\lambda}\int\psi(\lambda,\theta)\,\mathrm{d}\rho(\theta)+\frac{\mathcal{K}(\rho,\nu)+\log(\epsilon^{-1})}{n\lambda}.$$

Preuve. D'après l'inégalité de Jensen, lorsque ρ vérifie les hypothèses,

$$\begin{aligned}\exp\biggl[\int_\Theta n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\,\mathrm{d}\rho(\theta)-\mathcal{K}(\rho,\nu)\biggr]\\ \leq\int_\Theta\exp\Bigl\{n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\Bigr\}\mathbb{1}\biggl(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta)>0\biggr)\biggl(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta)\biggr)^{-1}\mathrm{d}\rho(\theta)\\ =\int_\Theta\exp\Bigl\{n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\Bigr\}\mathbb{1}\biggl(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta)>0\biggr)\,\mathrm{d}\nu(\theta)\\ \leq\int_\Theta\exp\Bigl\{n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\Bigr\}\,\mathrm{d}\nu(\theta).\end{aligned}$$

On peut ensuite appliquer le théorème de Fubini pour les fonctions positives.

$$\begin{aligned}\mathbb{E}\biggl\{\exp\biggl[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\int_\Theta n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\,\mathrm{d}\rho(\theta)-\mathcal{K}(\rho,\nu)\biggr]\biggr\}\\ \leq\mathbb{E}\biggl[\int_\Theta\exp\Bigl\{n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\Bigr\}\,\mathrm{d}\nu(\theta)\biggr]\\ =\int_\Theta\mathbb{E}\biggl[\exp\Bigl\{n\bigl[\lambda M(\theta)-\psi(\lambda,\theta)\bigr]\Bigr\}\biggr]\,\mathrm{d}\nu(\theta)=1.\end{aligned}$$

L'espérance indiquée dans la proposition ne porte pas forcément sur une fonction mesurable, mais cette fonction est majorée par une fonction mesurable d'espérance inférieure ou égale à 1 : c'est ce que montre la preuve et c'est le sens technique à donner à la proposition. La deuxième partie de la proposition s'obtient en appliquant l'inégalité de Markov. Là encore, l'événement en question n'est pas forcément mesurable : il faut comprendre qu'il contient un événement mesurable de probabilité au moins égale à $1-\epsilon$. □

Posons $m_i(\theta)=\mathbb{E}\bigl[f(X_i,\theta)\bigr]$,

$$\begin{aligned} v(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \Bigl\{ \bigl[f(X_i, \theta) - m_i(\theta) \bigr]^2 \Bigr\}, \\ \overline{V}(\lambda, \theta) &= \frac{2}{\lambda^2} \bigl[\psi(\lambda, \theta) - \lambda m(\theta) \bigr], \\ V(\lambda, \theta) &= \sup_{\beta \in [0, \lambda]} \overline{V}(\beta, \theta) \end{aligned}$$

et supposons que $v \overset{\text{def}}{=} \sup_{\theta \in \Theta} v(\theta) < \infty$ et $V(\lambda) \overset{\text{def}}{=} \sup_{\theta \in \Theta} V(\lambda, \theta) < \infty$, $0 \leq \lambda < \Lambda'$.

PROPOSITION 2.2 *Sous les hypothèses précédentes, pour toute constante positive c,*

$$\mathbb{E} \Biggl(\sup \biggl\{ \int_\Theta \bigl[M(\theta) - m(\theta) \bigr] \, \mathrm{d}\rho(\theta) ; \\ \rho \in \mathcal{M}_+^1(\Theta), \theta \mapsto M(\theta) - m(\theta) \in \mathbb{L}^1(\rho), \mathcal{K}(\rho, \nu) \leq c \biggr\} \Biggr) \\ \leq \inf_{\lambda \in [0, \Lambda'[} \frac{\lambda \overline{V}(\lambda)}{2} + \frac{c}{\lambda n} \leq \sqrt{\frac{2c}{n} V \biggl(\sqrt{\frac{2c}{nv}} \biggr)}.$$

En particulier quand Θ est fini, en prenant $c = \log(|\Theta|)$, $\rho = \delta_\theta$ et $\nu(\theta) = |\Theta|^{-1}$, $\theta \in \Theta$, on obtient

$$\mathbb{E} \Bigl\{ \sup_{\theta \in \Theta} \bigl[M(\theta) - m(\theta) \bigr] \Bigr\} \leq \sqrt{\frac{2 \log\bigl(|\Theta|\bigr)}{n} V \Biggl(\sqrt{\frac{2 \log\bigl(|\Theta|\bigr)}{nv}} \Biggr)}.$$

PREUVE. D'après la preuve de la proposition précédente, l'argument de l'espérance à majorer est inférieur ou égal à

$$\frac{1}{n\lambda} \log \biggl\{ \int \exp \Bigl[n \bigl[\lambda M(\theta) - \psi(\lambda, \theta) \bigr] \Bigr] \, \mathrm{d}\nu(\theta) \biggr\} + \frac{\lambda \overline{V}(\lambda)}{2} + \frac{c}{\lambda n},$$

et on conclut à l'aide de l'inégalité de Jensen. On otient ainsi le premier majorant $\displaystyle\inf_{\lambda \in [0, \Lambda'[} \frac{\lambda \overline{V}(\lambda)}{2} + \frac{c}{\lambda n}$ que l'on peut affaiblir en $\displaystyle\inf_{0 \leq \lambda \leq \beta} \frac{\lambda V(\beta)}{2} + \frac{c}{\lambda n}$. On obtient le second majorant en choisissant $\beta = \sqrt{\dfrac{2c}{nv}}$ et $\lambda = \sqrt{\dfrac{2c}{nV(\beta)}} \leq \beta$.

□

PROPOSITION 2.3 *Sous les hypothèses précédentes, pour toute constante positive c, avec probabilité au moins $1-\epsilon$,*

$$\sup\biggl\{ \int_{\Theta} \bigl[M(\theta) - m(\theta) \bigr] \, \mathrm{d}\rho(\theta); \\ \rho \in \mathcal{M}_+^1(\Theta), \theta \mapsto M(\theta) - m(\theta) \in \mathbb{L}^1(\Theta), \mathcal{K}(\rho, \nu) \leq c \biggr\} \\ \leq \inf_{\lambda \in [0, \Lambda'[} \frac{\lambda \overline{V}(\lambda)}{2} + \frac{c + \log(\epsilon^{-1})}{\lambda n} \leq \sqrt{\frac{2\bigl[c + \log(\epsilon^{-1})\bigr]}{n} V\Biggl(\sqrt{\frac{2\bigl[c+\log(\epsilon^{-1})\bigr]}{nv}}\Biggr)}.$$

En particulier quand Θ est fini, avec probabilité au moins $1-\epsilon$

$$\sup_{\theta \in \Theta} \bigl[M(\theta) - m(\theta) \bigr] \leq \sqrt{\frac{2 \log\bigl(|\Theta|/\epsilon\bigr)}{n} V\Biggl(\sqrt{\frac{2\log\bigl(|\Theta|/\epsilon\bigr)}{nv}} \Biggr)}.$$

PREUVE. C'est une conséquence directe de la seconde partie de la proposition 2.1 (page 7) et de la majoration $\psi(\lambda, \theta) \leq \dfrac{\lambda^2 V(\lambda)}{2} + \lambda m(\theta)$. □

PROPOSITION 2.4 *Supposons que $\Theta = \mathbb{B}_d = \bigl\{ \theta \in \mathbb{R}^d; \|\theta\| \leq 1 \bigr\}$ et qu'il existe deux constantes positives B et g telles que*

$$\begin{aligned} &\sup_{x \in \mathcal{X}} f(x, \theta) - \inf_{x \in \mathcal{X}} f(x, \theta) \leq B, && \theta \in \mathbb{B}_d, \\ &|f(x, \theta) - f(x, \theta')| \leq g \|\theta - \theta'\|, && x \in \mathcal{X}, \quad \theta, \theta' \in \mathbb{B}_d. \end{aligned}$$

Considérons le minimiseur du risque empirique

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{B}_d} M(\theta).$$

Avec probabilité au moins $1-\epsilon$,

$$m(\widehat{\theta}) \leq \inf_{\theta \in \mathbb{B}_d} m(\theta) + B \left\{ \sqrt{\frac{d}{2n} \log\biggl(1 + \frac{4g}{B} \sqrt{\frac{2n}{d}} \biggr) + \frac{\log\bigl(2/\epsilon\bigr)}{2n}} \right. \\ \left. + \sqrt{\frac{d}{8n}} + \sqrt{\frac{\log\bigl(2/\epsilon\bigr)}{2n}} \ \right\}.$$

Sous ces hypothèses très simples, on voit que la qualité de l'estimation de $\inf_{\theta \in \Theta} m(\theta)$ par $\widehat{\theta}$ dépend de la dimension d de l'espace des paramètres, et plus précisément du rapport d/n entre cette dimension et la taille de l'échantillon.

Preuve. Commençons par étendre le domaine de définition de f à $\mathbb{R}^d$ en posant

$$f(x, \theta) = f(x, \theta / \|\theta\|), \qquad \theta \in \mathbb{R}^d \setminus \mathbb{B}_d.$$

Soit $\delta > 0$ un paramètre réel positif dont nous choisirons par la suite la valeur et ν la mesure uniforme sur la boule $(1+\delta)\mathbb{B}_d$ de rayon $1+\delta$. Considérons pour tout $\theta \in \mathbb{B}_d$ la probabilité uniforme ρ_θ sur la boule $\theta + \delta \mathbb{B}_d$ centrée en θ de rayon θ. Le volume d'une boule de $\mathbb{R}^d$ étant proportionnel à son rayon élevé à la puissance d, on voit immédiatement que

$$\mathcal{K}(\rho_\theta, \nu) = d \log \left(\frac{1+\delta}{\delta} \right), \qquad \theta \in \mathbb{B}_d.$$

D'après la proposition précédente et l'inégalité de Hoeffding, avec probabilité au moins $1-\epsilon$, pour tout $\theta \in \mathbb{B}_d$,

$$\int m(\theta') \, \mathrm{d}\rho_\theta(\theta') \leq \int M(\theta') \, \mathrm{d}\rho_\theta(\theta') + B \sqrt{\frac{d \log\bigl(1 + \delta^{-1}\bigr) + \log\bigl(\epsilon^{-1}\bigr)}{2n}}.$$

On en déduit, toujours avec probabilité $1-\epsilon$,

$$m(\widehat{\theta}) \leq M(\widehat{\theta}) + 2g\delta + B \sqrt{\frac{d \log\bigl(1 + \delta^{-1}\bigr) + \log\bigl(\epsilon^{-1}\bigr)}{2n}}.$$

Soit $\theta_* \in \mathbb{B}_d$, tel que $m(\theta_*) = \inf_{\theta \in \mathbb{B}_d} m(\theta)$ (qui existe car $\theta \mapsto m(\theta)$ est continue sur le compact $\mathbb{B}_d$). Avec probabilité $1-\epsilon$

$$M(\theta_*) \leq m(\theta_*) + B \sqrt{\frac{\log\bigl(\epsilon^{-1}\bigr)}{2n}}.$$

Par construction de l'estimateur $\widehat{\theta}$, $M(\widehat{\theta}) \leq M(\theta_*)$. On en déduit donc qu'avec probabilité au moins $1-2\epsilon$,

$$m(\widehat{\theta}) \leq m(\theta_*) + B \left\{ \sqrt{\frac{d \log(1 + \delta^{-1}) + \log\bigl(\epsilon^{-1}\bigr)}{2n}} + \sqrt{\frac{\log(\epsilon^{-1})}{2n}} \right\} + 2g\delta.$$

On conclut en choisissant $\delta = \dfrac{B}{4g} \sqrt{\dfrac{d}{2n}}$ et en remplaçant ϵ par $\epsilon/2$. □

Proposition 2.5 *Supposons que* $\Theta = \mathbb{R}^d$*, qu'il existe une fonction mesurable* $(x,\theta) \mapsto \nabla f(x,\theta) \in \mathbb{R}^d$ *et des constantes positives* g *et* H *telles que*

$$|f(x,\theta) - f(x,\theta')| \leq g\|\theta - \theta'\|,$$
$$|f(x,\theta') - f(x,\theta) - \langle \nabla f(x,\theta), \theta' - \theta\rangle| \leq \frac{H}{2}\|\theta' - \theta\|^2, \quad x \in \mathcal{X}, \quad \theta, \theta' \in \mathbb{R}^d.$$

Soit $\theta_* \in \arg\min_{\theta \in \mathbb{B}_d} m(\theta)$*. Introduisons la fonction*

$$\chi(h) = \sup_{\theta \in \mathbb{B}_d} \frac{h}{2}\|\theta - \theta_*\|^2 - m(\theta) + m(\theta_*),$$

Dans ces conditions, le minimiseur empirique $\widehat{\theta} \in \arg\min_{\theta \in \mathbb{B}_d} M(\theta)$ *de* m *sur la boule unité vérifie avec probabilité au moins* $1 - \epsilon$

$$\|\widehat{\theta} - \theta_*\|^2 \leq \frac{8g^2}{nh^2}\left[\left(\frac{8H}{h} + 1\right)d + 2\log\bigl(\epsilon^{-1}\bigr)\right] + \frac{4\chi(h)}{h}$$
$$\text{et } m(\widehat{\theta}) - m(\theta_*) \leq \frac{4g^2}{nh}\left[\left(\frac{8H}{h} + 1\right)d + 2\log\bigl(\epsilon^{-1}\bigr)\right] + \chi(h).$$

Dans le cas où il existe $h > 0$ tel que $\chi(h) = 0$, on obtient donc une vitesse de convergence en d/n au lieu d'une vitesse en $\sqrt{d/n}$ sous les hypothèses plus faibles de la proposition précédente.

Exercice 1 *Dans le cas où* $m(\theta) - m(\theta_*) \geq c\, \|\theta - \theta_*\|^\alpha$*,* $\theta \in \mathbb{B}_d$*, avec* $c > 0$ *et* $\alpha > 2$ *quelle vitesse obtient-on ?*

Preuve. Choisissons $\rho_\theta = \mathcal{N}(\theta, \beta^{-1}I)$ et $\nu = \rho_{\theta_*}$. Remarquons que $\mathcal{K}(\rho_\theta, \nu) = \frac{\beta}{2}\|\theta - \theta_*\|^2$. Nous allons appliquer la proposition 2.1 (page 7) à la fonction $(x,\theta) \mapsto f(x,\theta_*) - f(x,\theta)$. D'après l'inégalité de Hoeffding

$$\log \mathbb{E}\exp\Bigl\{\lambda\bigl[f(X,\theta_*) - f(X,\theta)\bigr]\Bigr\} - \lambda\bigl[m(\theta_*) - m(\theta)\bigr] \leq \frac{\lambda^2 g^2\|\theta - \theta_*\|^2}{2}$$

Avec probabilité au moins $1 - \epsilon$, pour tout $\theta \in \mathbb{B}_d$,

$$\int m(\theta')\, \mathrm{d}\rho_\theta(\theta') - m(\theta_*) \leq \int M(\theta')\, \mathrm{d}\rho_\theta(\theta') - M(\theta_*)$$
$$+ \frac{\lambda g^2}{2}\int \|\theta' - \theta_*\|^2\, \mathrm{d}\rho_\theta(\theta') + \frac{\beta\|\theta - \theta_*\|^2}{2n\lambda} + \frac{\log(\epsilon^{-1})}{n\lambda}.$$

De plus

$$\int m(\theta') \,\mathrm{d}\rho_\theta(\theta') = m(\theta) \\ + \mathbb{E}\biggl[\int \Bigl[f(X,\theta') - f(X,\theta) - \langle \nabla f(X,\theta), \theta' - \theta \rangle \Bigr] \,\mathrm{d}\rho_\theta(\theta') \\ \geq m(\theta) - \frac{H}{2} \int \lVert \theta' - \theta \rVert^2 \,\mathrm{d}\rho_\theta(\theta') = m(\theta) - \frac{Hd}{2\beta}.$$

De même $\displaystyle\int M(\theta')\,\mathrm{d}\rho_\theta(\theta') \leq M(\theta) + \frac{Hd}{2\beta}$. On en déduit qu'avec probabilité au moins $1-\epsilon$, pour tout $\theta \in \mathbb{B}_d$,

$$m(\theta) - m(\theta_*) \leq M(\theta) - M(\theta_*) + \frac{Hd}{\beta} + \frac{\lambda g^2 d}{2\beta} + \frac{\lambda g^2}{2} \lVert \theta - \theta_* \rVert^2 \\ + \frac{\beta \lVert \theta - \theta_* \rVert^2}{2n\lambda} + \frac{\log\bigl(\epsilon^{-1}\bigr)}{n\lambda}.$$

On peut alors utiliser le fait que $m(\theta) - m(\theta_*) \geq \dfrac{h}{2} \lVert \theta - \theta_* \rVert^2 - \chi(h)$ et que par construction $M(\widehat{\theta}) \leq M(\theta_*)$. On en conclut avec probabilité au moins $1-\epsilon$

$$\frac{h}{2} \lVert \widehat{\theta} - \theta_* \rVert^2 \leq \chi(h) + \frac{d}{\beta}\biggl(H + \frac{\lambda g^2}{2} \biggr) \\ + \biggl(\frac{\lambda g^2}{2} + \frac{\beta}{2n\lambda} \biggr) \lVert \widehat{\theta} - \theta_* \rVert^2 + \frac{\log\bigl(\epsilon^{-1}\bigr)}{n\lambda}.$$

Ainsi

$$\lVert \widehat{\theta} - \theta_* \rVert^2 \biggl(1 - \frac{\lambda g^2}{h} - \frac{\beta}{n\lambda h} \biggr) \leq \frac{2\chi(h)}{h} + \frac{2d}{\beta h} \biggl(H + \frac{\lambda g^2}{2} \biggr) + \frac{2\log\bigl(\epsilon^{-1}\bigr)}{hn\lambda}.$$

Choisissons alors $\lambda = \dfrac{h}{4g^2}$ et $\beta = \dfrac{n\lambda h}{4} = \dfrac{nh^2}{16g^2}$. On obtient

$$\frac{1}{2} \lVert \widehat{\theta} - \theta_* \rVert^2 \leq \frac{2\chi(h)}{h} + \frac{32 g^2 d}{nh^3} \biggl(H + \frac{h}{8} \biggr) + \frac{8 g^2 \log\bigl(\epsilon^{-1}\bigr)}{nh^2},$$

qui donne la première majoration de la proposition.

Pour prouver la seconde, on utilise $\lVert \widehat{\theta} - \theta_* \rVert^2 \leq \dfrac{2}{h} \bigl[m(\widehat{\theta}\,) - m(\theta_*) + \chi(h) \bigr]$, pour obtenir

$$m(\widehat{\theta}\,) - m(\theta_*) \leq \frac{d}{\beta}\biggl(H + \frac{\lambda g^2}{2} \biggr) \\ + \biggl(\frac{\lambda g^2}{2} + \frac{\beta}{2n\lambda} \biggr) \frac{2}{h} \bigl[m(\widehat{\theta}\,) - m(\theta_*) + \chi(h) \bigr] + \frac{\log\bigl(\epsilon^{-1}\bigr)}{n\lambda}.$$

On conclut de même en remplaçant λ et β par leurs valeurs. $\square$

3. PAC-Bayes bounds for supervised classification

In this section, we are given some i.i.d. sample $(W_i)_{i=1}^n \in \mathcal{W}^n$, where $\mathcal{W}$ is a measurable space, and some binary measurable loss function $L : \mathcal{W} \times \Theta \rightarrow \{0, 1\}$, where Θ is a measurable parameter space. Our aim is to minimize with respect to $\theta \in \Theta$ the expected loss

$$\int L(w, \theta) \,\mathrm{d}\mathbb{P}(w),$$

where $\mathbb{P}$ is the marginal distribution of the observed sample $(W_i)_{i=1}^n$. More precisely, assuming that $\mathbb{P}$ is unknown, we would like to find an estimator $\widehat{\theta}(W_1^n)$ depending on the observed sample W_1^n such that the excess risk

$$\int L(w, \widehat{\theta}) \,\mathrm{d}\mathbb{P}(w) - \inf_{\theta \in \Theta} \int L(w, \theta) \,\mathrm{d}\mathbb{P}(w)$$

is small. The previous quantity is random, since $\widehat{\theta}$ depends on the random sample W_1^n. Therefore, how small it is can be understood in different ways. Here we will focus on the *deviations* of the excess risk. Accordingly, we will look for estimators providing a small risk with a probability close to one.

A typical example of such a problem is provided by supervised classification. In this setting $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is a finite set, $W_i = (X_i, Y_i)$, where (X_i, Y_i) are input-output pairs, a family of measurable classification rules $\bigl\{ f_\theta : \mathcal{X} \rightarrow \mathcal{Y}; \theta \in \Theta \bigr\}$ is considered and the loss function $L(w, \theta)$ is defined as the classification error

$$L\bigl[(x, y), \theta\bigr] = \mathbb{1}\bigl[f_\theta(x) \neq y\bigr].$$

Accordingly the aim is to minimize the expected classification error

$$\mathbb{P}_{X,Y}\bigl[f_\theta(X) \neq Y\bigr]$$

in view of a sample $(X_i, Y_i)_{i=1}^n$ of observations.

3.1. DEVIATION BOUNDS FOR SUMS OF BERNOULLI RANDOM VARIABLES. Given some parameter $\lambda \in \mathbb{R}$, let us consider the (normalized) log-Laplace transform of the Bernoulli distribution :

$$\Phi_\lambda(p) \overset{\text{def}}{=} -\frac{1}{\lambda} \log\bigl[1 - p + p \exp(-\lambda)\bigr].$$

Let us also consider the Kullback-Leibler divergence of two Bernoulli distributions

$$K(q,p) \overset{\text{def}}{=} q \log\left(\frac{q}{p}\right) + (1-q) \log\left(\frac{1-q}{1-p}\right).$$

In the sequel $\overline{\mathbb{P}}$ will be the empirical measure

$$\overline{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{W_i}$$

of an i.i.d. sample $(W_i)_{i=1}^n$ drawn from $\mathbb{P}^{\otimes n} \in \mathcal{M}_+^1(\mathcal{W}^n)$. We will use a short notation for integrals, putting for any $\rho, \pi \in \mathcal{M}_+^1(\Theta)$ and any integrable function $f \in \mathbb{L}_1\bigl(\mathcal{W} \times \Theta^2, \mathbb{P} \otimes \pi \otimes \rho\bigr)$

$$f(\mathbb{P}, \rho, \pi) = \int f(w, \theta, \theta') \,\mathrm{d}\mathbb{P}(w) \mathrm{d}\rho(\theta) \mathrm{d}\pi(\theta'),$$

so that for instance $L(\mathbb{P}, \rho) = \displaystyle\int L(w, \theta) \mathrm{d}\mathbb{P}(w) \mathrm{d}\rho(\theta)$.

Let us recall first Chernoff's bound.

PROPOSITION 3.1 *For any fixed value of the parameter $\theta \in \Theta$, the identity*

$$\int \exp\bigl[-n\lambda L(\overline{\mathbb{P}}, \theta)\bigr] \,\mathrm{d}\mathbb{P}^{\otimes n} = \exp\Bigl\{-n\lambda \Phi_\lambda\bigl[L(\mathbb{P}, \theta)\bigr]\Bigr\}$$

shows that with probability at least $1 - \epsilon$,

$$\begin{aligned} & L(\mathbb{P}, \theta) \leq B_+\bigl[L(\overline{\mathbb{P}}, \theta), \log(\epsilon^{-1})/n\bigr], \\ \textit{where} \quad & B_+(q, \delta) = \inf_{\lambda \in \mathbb{R}_+} \Phi_\lambda^{-1}\left(q + \frac{\delta}{\lambda}\right) \\ & \qquad = \sup\Bigl\{p \in [0,1] \,:\, K(q,p) \leq \delta\Bigr\}, \qquad q \in [0,1],\ \delta \in \mathbb{R}_+. \end{aligned}$$

Moreover

$$-\delta q \leq B_+(q, \delta) - q - \sqrt{2\delta q(1-q)} \leq 2\delta(1-q).$$

In the same way, the identity

$$\int \exp\bigl[n\lambda L(\overline{\mathbb{P}},\theta)\bigr] d\mathbb{P}^{\otimes n} = \exp\Bigl\{n\lambda \Phi_{-\lambda}\bigl[L(\mathbb{P},\theta)\bigr]\Bigr\}$$

shows that with probability at least $1-\epsilon$

$$\begin{aligned} & L(\overline{\mathbb{P}},\theta) \leq B_{-}\bigl[L(\mathbb{P},\theta), \log(\epsilon^{-1})/n\bigr], \\ \textit{where} \quad & B_{-}(q,\delta) = \inf_{\lambda \in \mathbb{R}_+} \Phi_{-\lambda}(q) + \frac{\delta}{\lambda} \\ & \qquad = \sup\Bigl\{p \in [0,1] \,:\, K(p,q) \leq \delta\Bigr\}, \qquad q \in [0,1],\ \delta \in \mathbb{R}_+, \end{aligned}$$

and

$$-\delta q \leq B_{-}(q,\delta) - q - \sqrt{2\delta q(1-q)} \leq 2\delta(1-q).$$

Let us mention here some important identity.

Proposition 3.2 *For any probability measures* π *and* ρ *on some measurable space, such that* $\mathcal{K}(\rho,\pi) < \infty$*, and any bounded measurable function* h*, let us define the transformed probability measure* $\pi_{\exp(h)} \ll \pi$ *by its density*

$$\frac{\mathrm{d}\pi_{\exp(h)}}{\mathrm{d}\pi} = \frac{\exp(h)}{Z},$$

where $Z = \int \exp(h)\,\mathrm{d}\pi$*. Let us moreover define*

$$\mathbf{Var}\bigl(h\,\mathrm{d}\pi\bigr) = \int \bigl(h - \int h\,\mathrm{d}\pi\bigr)^2\,\mathrm{d}\pi.$$

The expectations with respect to ρ *and* π *of* h *and the* log*-Laplace transform of* h *are linked by the identities*

$$\int h\,\mathrm{d}\rho - \mathcal{K}(\rho,\pi) + \mathcal{K}(\rho,\pi_{\exp(h)}) = \log\bigl[\int \exp(h)\,\mathrm{d}\pi\bigr] \tag{1}$$

$$= \int h\,\mathrm{d}\pi + \int_0^1 (1-\alpha)\,\mathbf{Var}\bigl[h\,\mathrm{d}\pi_{\exp(\alpha h)}\bigr]\,\mathrm{d}\alpha. \tag{2}$$

Proof. The first identity is a straightforward consequence of the definitions of $\pi_{\exp(h)}$ and of the Kullback-Leibler divergence function. The second one is the Taylor expansion of order one with integral remainder of the function

$$f(\alpha) = \log\bigl[\int \exp(\alpha h)\,\mathrm{d}\pi\bigr],$$

which says that $f(1) = f(0) + f'(0) + \int_0^1 (1-\alpha) f''(\alpha)\,\mathrm{d}\alpha$. □

Exercise 1 *Prove that $f \in \mathcal{C}^\infty$. Hint : write*

$$h^k \exp(\alpha h) = h^k + \int_0^{+\infty} \mathbb{1}(\gamma \leq \alpha) h^{k+1} \exp(\gamma h) \,\mathrm{d}\gamma$$

and use Fubini's theorem to show that $\alpha \mapsto \int h^k \exp(\alpha h) \,\mathrm{d}\pi$ belongs to $\mathcal{C}^1$ and compute its derivative.

Let us come now to the proof of Proposition 3.1 (page 15). Chernoff's inequality reads

$$\Phi_\lambda\bigl[L(\mathbb{P}, \theta)\bigr] - \frac{\log(\epsilon^{-1})}{n\lambda} \leq L(\overline{\mathbb{P}}, \theta),$$

where the inequality holds with probability at least $1 - \epsilon$. Since the left-hand side is non-random, it can be optimized in λ, giving

$$L(\mathbb{P}, \theta) \leq B_+\bigl[L(\overline{\mathbb{P}}, \theta), \log(\epsilon^{-1})/n\bigr].$$

Exercise 2 *Prove this statement in more details. For any integer $k > 1$, consider the event*

$$A_k = \Bigl\{ \sup_{\lambda \in \mathbb{R}_+} F(\lambda) - k^{-1} > L(\overline{\mathbb{P}}, \theta) \Bigr\},$$

where $F(\lambda) = \Phi_\lambda\bigl[L(\mathbb{P}, \theta)\bigr] - \dfrac{\log(\epsilon^{-1})}{n\lambda}$. Show that $\mathbb{P}^{\otimes n}(A_k) \leq \epsilon$ by choosing some suitable value of λ. Remark that $A_k \subset A_{k+1}$ and conclude that $\mathbb{P}^{\otimes n}\bigl(\cup_k A_k\bigr) \leq \epsilon$.

Since

$$\lim_{\lambda \to +\infty} \Phi_\lambda^{-1}\left(q + \frac{\delta}{\lambda}\right) = \lim_{\lambda \to +\infty} \frac{1 - \exp(-\lambda q - \delta)}{1 - \exp(-\lambda)} \leq 1,$$

$B_+(q, \delta) \leq 1$.

Applying equation (1, page 16) to Bernoulli distributions gives

$$\lambda \Phi_\lambda(p) = \lambda q + K(q, p) - K(q, p_\lambda)$$

where

$$p_\lambda = \frac{p}{p + (1 - p)\exp(\lambda)}.$$

This shows that

$$\begin{aligned} B_+(q, \delta) &= \sup\Bigl\{ p \in [0, 1] \;:\; \Phi_\lambda(p) \leq q + \frac{\delta}{\lambda},\ \lambda \in \mathbb{R}_+ \Bigr\} \\ &= \sup\Bigl\{ p \in [q, 1[\,:\; K(q, p) \leq \delta + K(q, p_\lambda), \lambda \in \mathbb{R}_+ \Bigr\} \end{aligned}$$

$$= \sup\Bigl\{ p \in [q,1[\, : \, K(q,p) \leq \delta \Bigr\}$$
$$= \sup\Bigl\{ p \in [0,1] \, : \, K(q,p) \leq \delta \Bigr\},$$

because when $q \leq p < 1$ then $\lambda = \log\left(\dfrac{q^{-1}-1}{p^{-1}-1}\right) \in \mathbb{R}_+$, $q = p_\lambda$ and therefore $K(q, p_\lambda) = 0$.

Let us remark now that $\dfrac{\partial^2}{\partial x^2} K(x,p) = x^{-1}(1-x)^{-1}$. Thus if $p \geq q \geq 1/2$, then

$$K(q,p) \geq \frac{(p-q)^2}{2q(1-q)},$$

so that if $K(q,p) \leq \delta$, then

$$p \leq q + \sqrt{2\delta q(1-q)}.$$

Now if $q \leq 1/2$ and $p \geq q$ then

$$K(q,p) \geq \left\{ \begin{matrix} \dfrac{(p-q)^2}{2p(1-p)}, & p \leq 1/2 \\ 2(p-q)^2, & p \geq 1/2 \end{matrix} \right\} \geq \frac{(p-q)^2}{2p(1-q)},$$

so that if $K(q,p) \leq \delta$, then

$$(p-q)^2 \leq 2\delta p(1-q),$$

implying that

$$p - q \leq \delta(1-q) + \sqrt{2\delta q(1-q) + \delta^2(1-q)^2} \leq \sqrt{2\delta q(1-q)} + 2\delta(1-q).$$

On the other hand,

$$K(q,p) \leq \frac{(p-q)^2}{2\min\{q(1-q), p(1-p)\}} \leq \frac{(p-q)^2}{2q(1-p)},$$

thus when $K(q,p) = \delta$ with $p > q$, then

$$(p-q)^2 \geq 2\delta q(1-p),$$

implying that

$$p - q \geq -\delta q + \sqrt{2\delta q(1-q) + \delta^2 q^2} \geq \sqrt{2\delta q(1-q)} - \delta q.$$

Exercise 3 *The second part of Proposition 3.1 (page 15) is proved in the same way and left as an exercise.*

3.2. PAC-Bayes bounds. We are now going to make Proposition 3.1 uniform with respect to θ. The PAC-Bayes approach to this is to randomize θ, so we will consider now joint distributions on $(W_1, \dots, W_n, \theta)$, where the distribution of $(W_1, \dots, W_n)$ is still $\mathbb{P}^{\otimes n}$ and the conditional distribution of θ given the sample is given by some transition probability kernel $\rho : \mathcal{W}^n \rightarrow \mathcal{M}_+^1(\Theta)$, called in this context a posterior distribution *. This posterior distribution ρ will be compared with a prior (meaning non-random) probability measure $\pi \in \mathcal{M}_+^1(\Theta)$.

Proposition 3.3 *Let us introduce the notation*

$$B_\Lambda(q, \delta) = \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1}\biggl(q + \frac{\delta}{\lambda} \biggr).$$

For any prior probability measure $\pi \in \mathcal{M}_+^1(\Theta)$ *and any* $\lambda \in \mathbb{R}_+$,

$$\int \exp \biggl[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} n\lambda \Bigl\{ \Phi_\lambda \bigl[L(\mathbb{P}, \rho) \bigr] - L(\overline{\mathbb{P}}, \rho) \Bigr\} - \mathcal{K}(\rho, \pi) \biggr] \mathrm{d}\mathbb{P}^{\otimes n} \leq 1, \qquad (3)$$

and therefore for any finite set $\Lambda \subset \mathbb{R}_+$, *with probability at least* $1 - \epsilon$, *for any* $\rho \in \mathcal{M}_+^1(\Theta)$,

$$L(\mathbb{P}, \rho) \leq B_\Lambda \Biggl(L(\overline{\mathbb{P}}, \rho), \frac{\mathcal{K}(\rho, \pi) + \log\bigl(|\Lambda|/\epsilon\bigr)}{n} \Biggr),$$

Proof. The exponential moment inequality (3) is a consequence of equation (1, page 16), showing that

$$\exp \biggl\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} n\lambda \int \Bigl\{ \Phi_\lambda \bigl[L(\mathbb{P}, \theta) \bigr] - L(\overline{\mathbb{P}}, \theta) \Bigr\} \mathrm{d}\rho(\theta) - \mathcal{K}(\rho, \pi) \biggr\} \\ \leq \int \exp \Bigl[n\lambda \Bigl\{ \Phi_\lambda \bigl[L(\mathbb{P}, \theta) \bigr] - L(\overline{\mathbb{P}}, \theta) \Bigr\} \Bigr] \mathrm{d}\pi(\theta),$$

and of the fact that Φ_λ is convex, showing that

$$\Phi_\lambda \bigl[L(\mathbb{P}, \rho) \bigr] \leq \int \Phi_\lambda \bigl[L(\mathbb{P}, \theta) \bigr] \mathrm{d}\rho(\theta).$$

The deviation inequality follows as usual. □

*. We will assume that ρ is a regular conditional probability kernel, meaning that for any measurable set A the map $(w_1, \dots, w_n) \mapsto \rho(w_1, \dots, w_n)(A)$ is assumed to be measurable. We will also assume that the σ-algebra we consider on Θ is generated by a countable family of subsets. See [1][page 50] for more details

We cannot take the infimum on $\lambda \in \mathbb{R}_+$ as in Proposition 3.1 (page 15), because we can no more cast our deviation inequality in such a way that λ appears on some non-random side of the inequality. Nevertheless, we can get a more explicit bound from some specific choice of the set Λ.

Proposition 3.4 *Let us define the least increasing upper bound of the variance of a Bernoulli distribution of parameter $p \in [0,1]$ as*

$$\overline{v}(p) = \begin{cases} p(1-p), & p \leq 1/2, \\ 1/4, & \textit{otherwise.} \end{cases}$$

Let us choose some positive integer parameter m and let us put

$$t = \frac{1}{4} \log \left(\frac{n}{8 \log\bigl[(m+1)/\epsilon\bigr]} \right).$$

With probability at least $1-\epsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$,

$$L(\mathbb{P}, \rho) \leq L(\overline{\mathbb{P}}, \rho) + B_m\bigl[L(\overline{\mathbb{P}}, \rho), \mathcal{K}(\rho, \pi), \epsilon\bigr],$$

where

$$\begin{aligned} B_m\bigl(q, e, \epsilon\bigr) = \max \Biggl\{ & \sqrt{\frac{2\overline{v}(q)\bigl\{e + \log\bigl[(m+1)/\epsilon\bigr]\bigr\}}{n}} \cosh\bigl(t/m\bigr) \\ & \qquad + \frac{2(1-q)\bigl\{e + \log\bigl[(m+1)/\epsilon\bigr]\bigr\}}{n} \cosh(t/m)^2, \\ & \qquad\qquad \frac{2\bigl\{e + \log\bigl[(m+1)/\epsilon\bigr]\bigr\}}{n} \Biggr\} \\ \leq & \sqrt{\frac{2\overline{v}(q)\bigl\{e + \log\bigl[(m+1)/\epsilon\bigr]\bigr\}}{n}} \cosh\bigl(t/m\bigr) \\ & \qquad + \frac{2\bigl\{e + \log\bigl[(m+1)/\epsilon\bigr]\bigr\}}{n} \cosh(t/m)^2. \end{aligned}$$

Moreover, as soon as $n \geq 5$,

$$B_{\lfloor \log(n)^2 \rfloor - 1}(q, e, \epsilon) \leq B(q, e, \epsilon) \overset{\text{def}}{=} \\ \sqrt{\frac{2\overline{v}(q)\bigl\{e + \log\bigl[\log(n)^2/\epsilon\bigr]\bigr\}}{n}} \cosh\bigl[\log(n)^{-1}\bigr]$$

$$+ \frac{2\bigl\{e + \log\bigl[\log(n)^2/\epsilon\bigr]\bigr\}}{n} \cosh\bigl[\log(n)^{-1}\bigr]^2, \quad (4)$$

so that with probability at least $1-\epsilon$*, for any* $\rho \in \mathcal{M}_+^1(\Theta)$*,*

$$L(\mathbb{P}, \rho) \leq L(\overline{\mathbb{P}}, \rho) + \sqrt{\frac{2\overline{v}\bigl[L(\overline{\mathbb{P}},\rho)\bigr]\Bigl\{\mathcal{K}(\rho,\pi) + \log\bigl[\log(n)^2/\epsilon\bigr]\Bigr\}}{n}} \cosh\bigl[\log(n)^{-1}\bigr] + \frac{2\Bigl\{\mathcal{K}(\rho,\pi) + \log\bigl[\log(n)^2/\epsilon\bigr]\Bigr\}}{n} \cosh\bigl[\log(n)^{-1}\bigr]^2.$$

Proof. Let us put

$$\begin{aligned} q &= L(\overline{\mathbb{P}}, \rho), \\ \delta &= \frac{\mathcal{K}(\rho,\pi) + \log\bigl[(m+1)/\epsilon\bigr]}{n}, \\ \lambda_{\min} &= \sqrt{\frac{8\log\bigl[(m+1)/\epsilon\bigr]}{n}}, \\ \Lambda &= \Bigl\{\lambda_{\min}^{1-k/m}, k = 0, \dots, m\Bigr\}, \\ p &= B_\Lambda(q,\delta) = \inf_{\lambda\in\Lambda} \Phi_\lambda^{-1}\biggl(q + \frac{\delta}{\lambda}\biggr), \\ \widehat{\lambda} &= \sqrt{\frac{2\delta}{\overline{v}(p)}}. \end{aligned}$$

According to equation (2, page 16) applied to Bernoulli distributions, for any $\lambda \in \Lambda$,

$$\Phi_\lambda(p) = p - \frac{1}{\lambda}\int_0^\lambda (\lambda - \alpha)p_\alpha(1-p_\alpha)\,\mathrm{d}\alpha \leq q + \frac{\delta}{\lambda}.$$

As moreover $p_\alpha \leq p$,

$$p - q \leq \inf_{\lambda\in\Lambda} \frac{\lambda\overline{v}(p)}{2} + \frac{\delta}{\lambda} = \inf_{\lambda\in\Lambda} \sqrt{2\delta\overline{v}(p)}\cosh\biggl[\log\biggl(\frac{\widehat{\lambda}}{\lambda}\biggr)\biggr].$$

As $\overline{v}(p) \leq 1/4$ and $\delta \geq \dfrac{\log\bigl[(m+1)/\epsilon\bigr]}{n}$,

$$\sqrt{\frac{2\delta}{\overline{v}(p)}} = \widehat{\lambda} \geq \lambda_{\min} = \sqrt{\frac{8\log\bigl[(m+1)/\epsilon\bigr]}{n}}.$$

Therefore either $\lambda_{\min} \leq \widehat{\lambda} \leq 1$, or $\widehat{\lambda} > 1$. Let us consider these two cases separately.

If $\lambda_{\min} = \min \Lambda \leq \widehat{\lambda} \leq \max \Lambda = 1$, then $\log\bigl(\widehat{\lambda}\bigr)$ is at distance at most t/m from some $\log\bigl(\lambda\bigr)$ where $\lambda \in \Lambda$, because $\log(\Lambda)$ is a grid with constant steps of size $2t/m$. Thus

$$p - q \leq \sqrt{2\delta \overline{v}(p)} \cosh\bigl(t/m\bigr).$$

If moreover $q \leq 1/2$, then $\overline{v}(p) \leq p(1-q)$, so that we obtain a quadratic inequality in p, whose solution is less than

$$p \leq q + \sqrt{2\delta q(1-q)} \cosh\bigl(t/m\bigr) + 2\delta(1-q)\cosh\bigl(t/m\bigr)^2.$$

If on the contrary $q \geq 1/2$, then $\overline{v}(p) = \overline{v}(q) = 1/4$ and

$$p \leq q + \sqrt{2\delta \overline{v}(q)} \cosh\bigl(t/m\bigr),$$

so that in both cases

$$p - q \leq \sqrt{2\delta \overline{v}(q)} \cosh(t/m) + 2\delta(1-q)\cosh\bigl(t/m\bigr)^2. \tag{5}$$

Let us consider now the case when $\widehat{\lambda} > 1$. In this case $\overline{v}(p) < 2\delta$, so that

$$p - q \leq \frac{\overline{v}(p)}{2} + \delta \leq 2\delta.$$

In conclusion, applying Proposition 3.3 (page 19) we see that with probability at least $1-\epsilon$, for any posterior distribution ρ,

$$L(\mathbb{P}, \rho) \leq p \leq q + \max \Bigl\{ 2\delta, \sqrt{2\delta \overline{v}(q)} \cosh\bigl(t/m\bigr) + 2\delta(1-q)\cosh\bigl(t/m\bigr)^2 \Bigr\},$$

which is precisely the statement to be proved.

In the special case when $m = \lfloor \log(n)^2 \rfloor - 1 \geq \log(n)^2 - 2$,

$$\frac{t}{m} \leq \frac{1}{4\bigl[\log(n)^2 - 2\bigr]} \log\left(\frac{n}{8 \log\bigl[\log(n)^2 - 1\bigr]} \right) \leq \log(n)^{-1}$$

as soon as the last inequality holds, that is as soon as $n \geq \exp(\sqrt{2}) \simeq 4.11$ to make $\log(n)^2 - 2$ positive and

$$3\log(n)^2 - 8 + \log(n) \log\Bigl\{ 8 \log\bigl[\log(n)^2 - 1\bigr] \Bigr\} \geq 0,$$

which holds true for any $n \geq 5$, as can be checked numerically. □

4. Linear classification and support vector machines

We are going in this section to consider more specifically the case of linear binary classification. In this setting $\mathcal{W} = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \{-1, +1\}$, $w = (x, y)$, where $x \in \mathbb{R}^d$ and $y \in \{-1, +1\}$, $\Theta = \mathbb{R}^d$, and

$$L(w, \theta) = \mathbb{1}\bigl[\langle \theta, x \rangle y \leq 0\bigr].$$

Although we will stick in this presentation to the case when $\mathcal{X}$ is a vector space of finite dimension, the results also apply to support vector machines, where the pattern space is some arbitrary space mapped to a Hilbert space $\mathcal{H}$ by some implicit mapping $\Psi : \mathcal{X} \to \mathcal{H}$, $\Theta = \mathcal{H}$ and $L(w, \theta) = \mathbb{1}\bigl(\langle \theta, \Psi(x) \rangle y \leq 0\bigr)$. It turns out that classification algorithms do not need to manipulate $\mathcal{H}$ itself, but only to compute scalar products of the form $k(x_1, x_2) = \langle \Psi(x_1), \Psi(x_2) \rangle$, defining a symmetric positive kernel k on the original pattern space $\mathcal{X}$. The converse is also true, any positive symmetric kernel k can be represented as a scalar product in some mapped Hilbert space (this is the Moore-Aronszajn theorem). Often used kernels on $\mathbb{R}^d$ are

$$\begin{aligned} k(x_1, x_2) &= \bigl(1 + \langle x_1, x_2 \rangle\bigr)^s, \text{ for which } \dim \mathcal{H} < \infty, \\ k(x_1, x_2) &= \exp\bigl(-\lVert x_1 - x_2 \rVert^2\bigr), \text{ for which } \dim \mathcal{H} = +\infty. \end{aligned}$$

In the following, we will work in $\mathbb{R}^d$, which covers only the case when $\dim \mathcal{H} < \infty$, but extensions would be possible.

Let us consider, after [5, 8] as prior probability measure π the centered Gaussian measure with covariance $\beta^{-1}\,\mathbf{Id}$, so that

$$\frac{\mathrm{d}\pi}{\mathrm{d}\theta}(\theta) = \left(\frac{\beta}{2\pi}\right)^{d/2} \exp\left(-\frac{\beta \lVert \theta \rVert^2}{2}\right).$$

Let us also consider the function

$$\begin{aligned} \varphi(x) &= \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \exp\bigl(-t^2/2\bigr)\,\mathrm{d}t, \qquad x \in \mathbb{R} \\ &\leq \min\left\{\frac{1}{x\sqrt{2\pi}}, \frac{1}{2}\right\} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbb{R}_+. \end{aligned}$$

Let π_θ be the measure π shifted by θ, defined by the identity

$$\int h(\theta')\,\mathrm{d}\pi_\theta(\theta') = \int h(\theta + \theta')\,\mathrm{d}\pi(\theta').$$

In this case

$$\mathcal{K}(\pi_\theta, \pi) = \frac{\beta}{2} \lVert \theta \rVert^2,$$

and

$$L(w, \pi_\theta) = \varphi\bigl[\sqrt{\beta} y \lVert x \rVert^{-1} \langle \theta, x \rangle\bigr].$$

Thus the randomized loss function has an explicit expression : randomization replaces the indicator function of the negative real line by a smooth approximation. As we are eventually interested in $L(w, \theta)$, we will shift things a little bit, considering along with the classification error function L some *error with margin*

$$M(w, \theta) = \mathbb{1}\bigl[y \lVert x \rVert^{-1} \langle \theta, x \rangle \leq 1\bigr].$$

Unlike $L(w, \theta)$ which is independent of the norm of θ, the margin error $M(w, \theta)$ depends on $\lVert \theta \rVert$, counting a classification error each time x is at distance less than $\lVert x \rVert / \lVert \theta \rVert$ from the boundary $\{x' : \langle \theta, x' \rangle = 0\}$, so that the error with margin region is the complement of the open cone $\bigl\{x \in \mathbb{R}^d \,; y \langle \theta, x \rangle > \lVert x \rVert \bigr\}$.

Let us compute the randomized margin error

$$M(w, \pi_\theta) = \varphi\Bigl\{ \sqrt{\beta}\bigl[y \lVert x \rVert^{-1} \langle \theta, x \rangle - 1\bigr] \Bigr\}.$$

It satisfies the inequality

$$M(w, \pi_\theta) \geq \varphi(-\sqrt{\beta}\,) L(w, \theta) = \bigl[1 - \varphi\bigl(\sqrt{\beta}\,\bigr)\bigr] L(w, \theta). \tag{6}$$

Applying previous results we obtain

Proposition 4.1 *With probability at least $1 - \epsilon$, for any $\theta \in \mathbb{R}^d$,*

$$L(\mathbb{P}, \theta) \leq \bigl[1 - \varphi(\sqrt{\beta})\bigr]^{-1} M(\mathbb{P}, \pi_\theta) \leq C_1(\theta),$$

where

$$C_1(\theta) = \bigl[1 - \varphi\bigl(\sqrt{\beta}\,\bigr)\bigr]^{-1} B\left(M(\overline{\mathbb{P}}, \pi_\theta), \frac{\beta \lVert \theta \rVert^2}{2}, \epsilon \right),$$

the bound B being defined by equation (4, page 21).

We can now minimize this empirical upper-bound to define an estimator. Let us consider some estimator $\widehat{\theta}$ such that

$$C_1(\widehat{\theta}) \leq \inf_{\theta \in \mathbb{R}^d} C_1(\theta) + \zeta.$$

Then for any fixed parameter $\theta_\star$, $C_1(\theta) \leq C_1(\theta_\star) + \zeta$. On the other hand, with probability at least $1 - \epsilon$

$$M(\overline{\mathbb{P}}, \pi_{\theta_\star}) \leq B_-\left(M(\mathbb{P}, \pi_{\theta_\star}), \frac{\log(\epsilon^{-1})}{n} \right).$$

Indeed

$$\int \exp\Bigl\{n\lambda\bigl[M(\overline{\mathbb{P}},\pi_{\theta_\star}) - \Phi_{-\lambda}\bigl[M(\mathbb{P},\pi_{\theta_\star})\bigr]\Bigr\}\,\mathrm{d}\mathbb{P}^{\otimes n} \\ \leq \int \exp\Bigl\{n\lambda \int \Bigl\{M(\overline{\mathbb{P}},\theta) - \Phi_{-\lambda}\bigl[M(\mathbb{P},\theta)\Bigr\}\,\mathrm{d}\pi_{\theta_\star}(\theta)\Bigr\}\,\mathrm{d}\mathbb{P}^{\otimes n} \leq 1,$$

because $p \mapsto -\Phi_{-\lambda}(p)$ is convex. As a consequence

Proposition 4.2 *With probability at least* $1-2\epsilon$*,*

$$L(\mathbb{P},\widehat{\theta}) \leq \\ \inf_{\theta_\star\in\Theta}\bigl[1-\varphi\bigl(\sqrt{\beta}\,\bigr)\bigr]^{-1} B\biggl(B_-\biggl(M(\mathbb{P},\pi_{\theta_\star}), \frac{\log(\epsilon^{-1})}{n}\biggr), \frac{\beta\|\theta_\star\|^2}{2}, \epsilon\biggr) + \zeta.$$

It is also possible to state a result in terms of empirical margins. Indeed

$$M(w,\pi_\theta) \leq M(w,\theta/2) + \varphi(\sqrt{\beta}).$$

Thus with probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$L(\mathbb{P},\theta) \leq C_2(\theta),$$

where

$$C_2(\theta) = \bigl[1-\varphi\bigl(\sqrt{\beta}\,\bigr)\bigr]^{-1} B\biggl(M(\overline{\mathbb{P}},\theta/2) + \varphi\bigl(\sqrt{\beta}\,\bigr), \frac{\beta\|\theta\|^2}{2}, \epsilon\biggr).$$

However, C_1 and C_2 are non-convex criterions, faster minimization algorithms are available for the usual SVN loss function, for which it is also possible to derive some generalization bound. Indeed, let us choose some positive radius R and let us put $\|x\|_R = \max\bigl\{R, \|x\|\bigr\}$, so that in the case when $\|x\| \leq R$, $\|x\|_R = R$.

$$M(w,\pi_\theta) = \varphi\bigl[\sqrt{\beta}\bigl(y\|x\|^{-1}\langle\theta,x\rangle - 1\bigr)\bigr] \leq \bigl(2 - y\|x\|_R^{-1}\langle\theta,x\rangle\bigr)_+ + \varphi(\sqrt{\beta}). \quad (7)$$

To check that this is true, consider the functions

$$f(z) = \varphi\bigl[\sqrt{\beta}\bigl(\|x\|^{-1}z - 1\bigr)\bigr], \\ g(z) = \bigl(2 - \|x\|_R^{-1}z\bigr)_+ + \varphi(\sqrt{\beta}), \qquad z \in \mathbb{R}.$$

Let us remark that they are both non increasing, that f is convex on the interval $z \in \bigl(\|x\|, \infty\bigl($ (because φ is convex on $\mathbb{R}_+$), and that $\sup f = \sup \varphi = 1$. Since $\|x\|_R \geq \|x\|$, for any $z \in]-\infty, \|x\|]$, $g(z) \geq 1 \geq f(z)$. Moreover,

$g(2\|x\|_R) = \varphi(\sqrt{\beta}) \geq \varphi\bigl[\sqrt{\beta}\bigl(2\|x\|^{-1}\|x\|_R - 1\bigr)\bigr] = f(z)$. Since on the interval $\bigl(\|x\|, 2\|x\|_R\bigr)$, the function g is linear, the function f is convex and g is not smaller than f at the two ends, this proves that g is not smaller than f on the whole interval. Finally, on the interval $z \in \bigl(2\|x\|_R, +\infty\bigl($, the function g is constant and the function f is decreasing, so that on this interval also g is not smaller than f, and this ends the proof of (7), since the three intervals on which $g \geq f$ cover the whole real line.

Using the upper bounds (7) and (6, page 24), and Proposition 3.3 (page 19), we obtain

PROPOSITION 4.3 *With probability at least* $1 - \epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$
\begin{aligned}
L(\mathbb{P}, \theta) \leq \bigl[1 - \varphi\bigl(\sqrt{\beta}\,\bigr)\bigr]^{-1} B_\Lambda \biggl(\int \bigl(2 - y\|x\|_R^{-1}\langle \theta, x\rangle\bigr)_+ d\overline{\mathbb{P}}(x,y) + \varphi(\sqrt{\beta}), \\
\frac{\beta\|\theta\|^2 + 2\log\bigl(|\Lambda|/\epsilon\bigr)}{2n} \biggr) \\
= \bigl[1 - \varphi\bigl(\sqrt{\beta}\bigr)\bigr]^{-1} \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1} \biggl[C_3(\lambda, \theta) + \varphi\bigl(\sqrt{\beta}\,\bigr) + \frac{\log\bigl(|\Lambda|/\epsilon\bigr)}{n\lambda} \biggr],
\end{aligned}
$$

where

$$
C_3(\lambda, \theta) = \int \bigl(2 - y\|x\|_R^{-1}\langle \theta, x\rangle\bigr)_+ \,\mathrm{d}\overline{\mathbb{P}}(x,y) + \frac{\beta\|\theta\|^2}{2n\lambda}.
$$

Let us assume now that the patterns x are in a ball, so that $\|x\| \leq R$ almost surely. In this case $\|x\|_R = R$ almost surely. Let us remark that $L(\mathbb{P}, \theta) = L(\mathbb{P}, 2R\,\theta)$, and let us make the previous result uniform in $\beta \in \Xi$. This leads to

PROPOSITION 4.4 *Let us assume that* $\|x\| \leq R$ *almost surely. With probability at least* $1 - \epsilon$, *for all* $\theta \in \mathbb{R}^d$,

$$
\begin{aligned}
L(\mathbb{P}, \theta) \leq \inf_{\beta \in \Xi} \bigl[1 - \varphi(\sqrt{\beta})\bigr]^{-1} \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1} \biggl[2C_4\bigl(\beta, \lambda, \theta\bigr) \\
+ \varphi(\sqrt{\beta}) + \frac{\log\bigl(|\Xi|\,|\Lambda|/\epsilon\bigr)}{n\lambda} \biggr],
\end{aligned}
$$

where

$$
C_4(\beta, \lambda, \theta) = \frac{1}{2}\, C_3(\lambda, 2R\,\theta) = \int \bigl(1 - y\langle \theta, x\rangle\bigr)_+ \,\mathrm{d}\overline{\mathbb{P}}(x,y) + \frac{\beta R^2\|\theta\|^2}{n\lambda},
$$

and

$$
\Phi_\lambda^{-1}(q) = \frac{1 - \exp(-\lambda q)}{1 - \exp(-\lambda)} \leq \frac{q}{1 - \dfrac{\lambda}{2}}.
$$

The loss function $C_4(\lambda,\theta)$ is the most employed learning criterion for support vector machines, and is called the box constraint. It is convex in θ. There are fast algorithms to compute $\inf_\theta C_4(\lambda,\theta)$ for any fixed values of λ and β. Here we get an empirical criterion which could be used to optimize also the values of λ and β, that is to optimize the strength of the regularizing factor $\dfrac{\beta R^2 \lVert\theta\rVert^2}{n\lambda}$.

Here $\lVert\theta\rVert^{-1}$ can be interpreted as the margin width, that is the minimal distance of x from the separating hyperplane $\{x' : \langle \theta, x'\rangle = 0\}$ beyond which the error term $\bigl(1-y\langle\theta,x\rangle\bigr)_+$ vanishes (for data x that are on the right side of the separating hyperplane). The speed of convergence depends on $R^2\lVert\theta\rVert^2/n$. For this reason, $R^2\lVert\theta\rVert^2$, the square of the ratio between the radius of the ball containing the data and the margin, plays the same role as the dimension d in Proposition 2.4 (page 10). The bound does not depend on d, showing that with separating hyperplanes and more generally Support Vector Machines, we can get low error rates while choosing to represent the data in a Reproducing Kernel Hilbert Space with a large, or even infinite, dimension.

Here, we considered only linear hyperplane and data centered around 0. Anyhow, this also covers affine hyperplane and data contained in a non necessarily centered ball, through a change of coordinates. More precisely, the previous proposition has the following corollary :

Corollary 4.5 *Assume that almost surely $\lVert x-c\rVert \leq R$, for some $c \in \mathbb{R}^d$ and $R \in \mathbb{R}_+$. With probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$, any $\gamma \in \mathbb{R}$ such that $\min_{i=1,\dots,n} \langle\theta,x_i\rangle \leq \gamma \leq \max_{i=1,\dots,n}\langle\theta,x_i\rangle$,*

$$\int \mathbb{1}\bigl[y\bigl(\langle\theta,x\rangle-\gamma\bigr)\leq 0\bigr]\,\mathrm{d}\mathbb{P}(x,y) \leq \inf_{\beta\in\Xi}\bigl[1-\varphi(\sqrt{\beta})\bigr]^{-1} \inf_{\lambda\in\Lambda}\Phi_\lambda^{-1}\left[2C_5(\beta,\lambda,\theta,\gamma)+\varphi(\sqrt{\beta})+\frac{\log\bigl(|\Xi|\,|\Lambda|/\epsilon\bigr)}{n\lambda}\right],$$

where

$$C_5(\beta,\lambda,\theta,\gamma) = \int\bigl[1-y\bigl(\langle\theta,x\rangle-\gamma\bigr)\bigr]_+\,\mathrm{d}\overline{\mathbb{P}}(x,y)+\frac{4\beta R^2\lVert\theta\rVert^2}{n\lambda}.$$

Proof. Let us apply the previous result to $x' = (x-c,R)$, and $\theta' = \bigl[\theta, R^{-1}\bigl(\langle\theta,c\rangle-\gamma\bigr)\bigr]$. We get that $\lVert x'\rVert^2 \leq 2R^2$ and $\lVert\theta'\rVert^2 = 2\lVert\theta\rVert^2$, because almost surely $-\lVert\theta\rVert R \leq \operatorname{ess\,inf}\langle\theta,x-c\rangle \leq \gamma - \langle\theta,c\rangle \leq \operatorname{ess\,sup}\langle\theta,x-c\rangle \leq \lVert\theta\rVert R$, so that almost surely, for the allowed values of γ, $\bigl(\langle\theta,c\rangle-\gamma\bigr)^2 \leq R^2\lVert\theta\rVert^2$. This proves that $C_4(\beta,\lambda,\theta') \leq C_5(\beta,\lambda,\theta,\gamma)$, as required to deduce the corollary from the previous proposition. □

Références

[1] O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI – 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004. Pages 1–269.

[2] O. Catoni. *PAC-Bayesian Supervised Classification : The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007. Pages i-xii, 1-163.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley and Sons, New York, second edition, 2006.

[4] Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 353–360, New York, NY, USA, 2009. ACM.

[5] J. Langford and J. Shawe-Taylor. PAC-bayes & margins. In *Advances in Neural Information Processing Systems*, pages 423–430, 2002.

[6] D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th annual conference on Computational Learning Theory*. Morgan Kaufmann, 1999.

[7] D. A. McAllester. PAC-Bayesian stochastic model selection. *Mach. Learn.*, 51(1) :5–21, April 2003.

[8] David Mcallester. Simplified pac-bayesian margin bounds. In *In COLT*, pages 203–215, 2003.

[9] M. Seeger. PAC-Bayesian generalization error bounds for gaussian process classification. Informatics report series EDI-INF-RR-0094, Division of Informatics, University of Edinburgh, 2002.